Information[a], Dynamics, and the Modeling of Images for the Purpose of Compression[b]

Andy Fraser
Portland State

CNLS-LANL Image Workshop 2002-12-6

[a] References to Cover and Thomas, “Elements of Information Theory”, 1991
[b] Lossless only

	Topic	Take home points
•	Simple examples	
•	Dynamical system	This can all be viewed as dynamics
•	McMillan Theorem	Entropy is growth rate
•	Typical set coding	$-\log \text{Probability} \equiv \text{Code Length}$
•	Relative entropy	$D(P\|\|Q) \equiv$ Cost of model error
•	Costs of block boundaries	Don't use blocks
•	AR models for image coding	High likelihood $\rightarrow$ compression

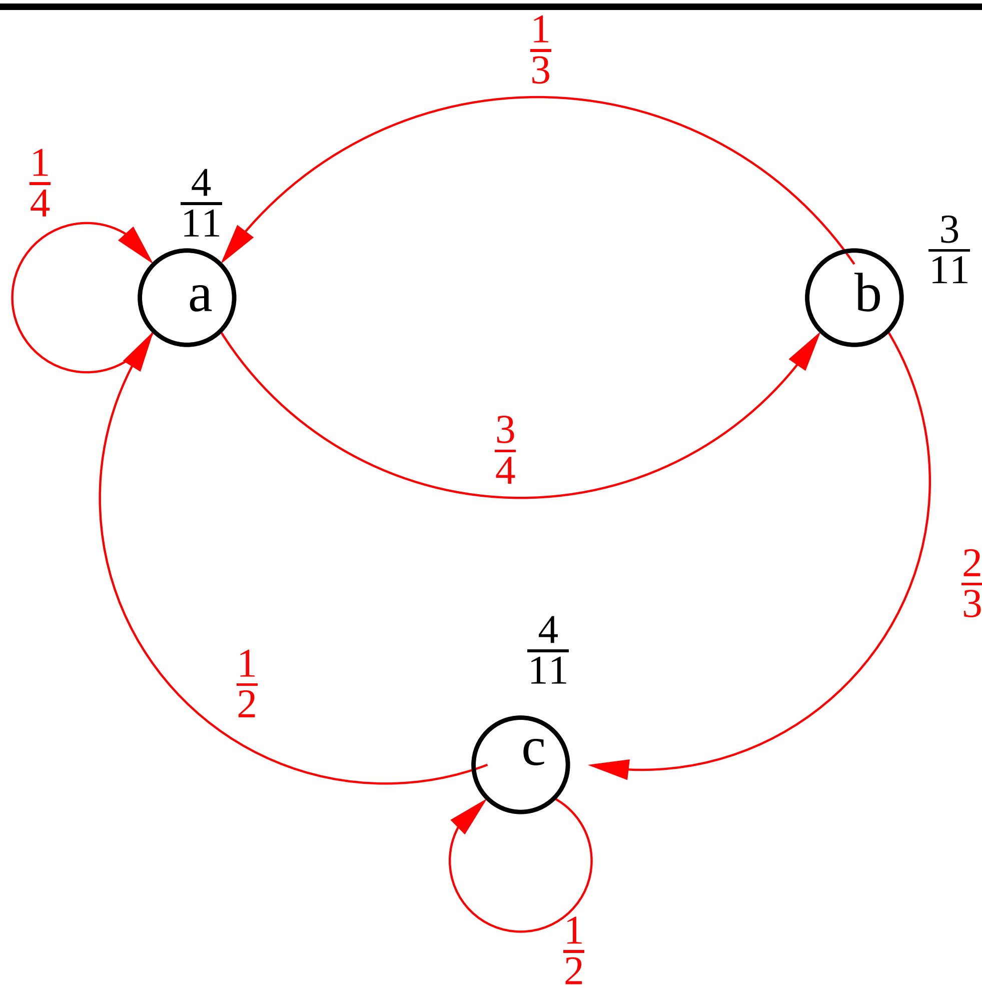
Compressing Strings of *Quits*

- Want to map string like $BACDAAAB\ldots A$ of i.i.d letters invertibly
- to short string like $01101010101110010\ldots 0$ of bits

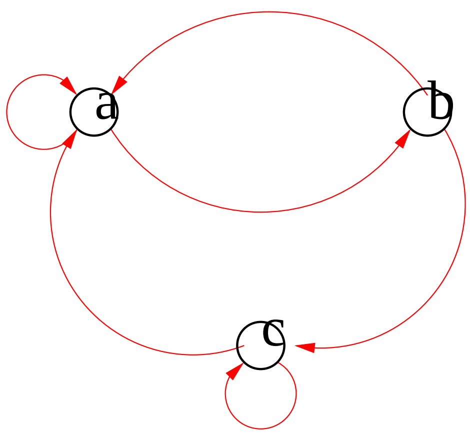
Letter	Code	Length L	Probability P	$P \cdot L$		
A	0	1	0.5	0.5		
B	10	2	0.25	0.5		
C	110	3	0.125	0.375		
D	111	3	0.125	0.375		
				1.75	$= \mathbb{E}L$	Optimal

because $-\log_2(P(x)) = L(x)\ \forall x$ and $\mathbb{E}L = H$

Markov *Trits*



Code for Trits

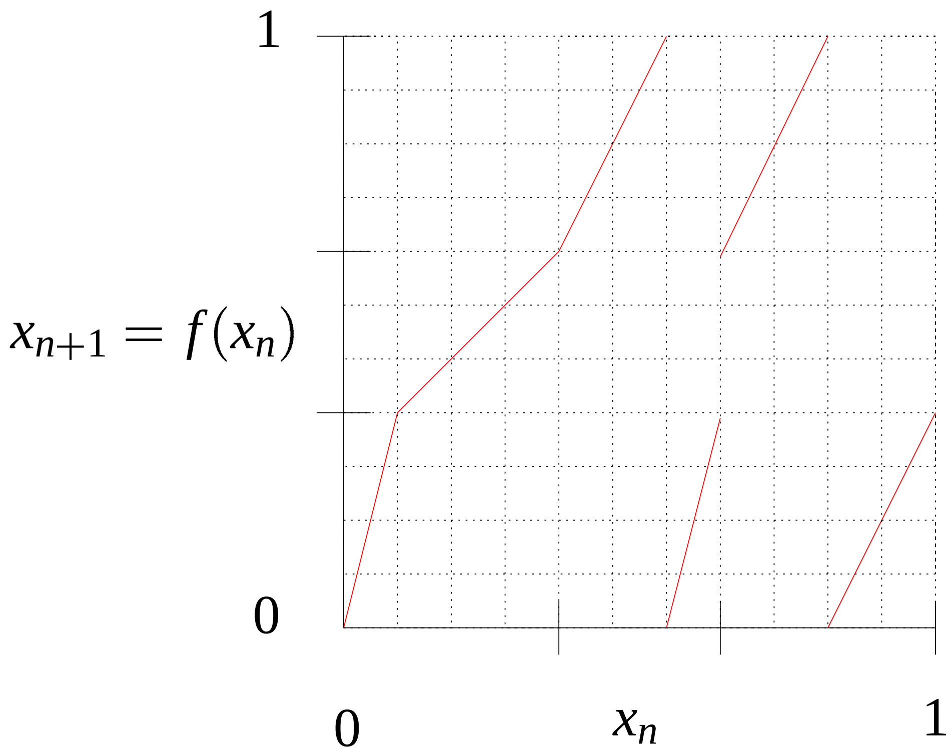


Given	Followed by	Code	Cond. Prob.
a	a	00	$\frac{1}{4}$
a	ba	01	$\frac{1}{4}$
a	bc	1	$\frac{1}{2}$
c	c	0	$\frac{1}{2}$
c	a	1	$\frac{1}{2}$

$$H = \frac{10}{11} \text{bits}$$

1-d Map

Avg. $\log_2(|f'|) = \frac{10}{11}$bits



Typical Sets and The McMillan Theorem

For large n, the number of plausible strings, x_1^n, is 2^{nH} and the probability of each plausible string is 2^{-nH} *excluding a set of strings whose total probability is small.*

Notation:

X A discrete random variable (carries alphabet and probability as freight).

$\mathcal{X}$ Alphabet, i.e., set of possible values.

P_X The function that maps subsets of $\mathcal{X}$ to probabilities.

$P(x)$, $\mathbf{P}r\{X = x\}$ Variants in notation.

x_1^n The sequence $(x_1, x_2, \ldots, x_n)$.

Definition 1 *The* typical set $A_\varepsilon^{(n)}$ *is the set of sequences such that*

$$2^{n(H(X)+\varepsilon)} \le P(x_1^n) \le 2^{n(H(X)-\varepsilon)}$$

Typical Sets[a] cont.

Theorem 1 *(McMillan) If the process $X_1, X_2, X_3, \ldots$ is ergodic with probabilities given by $\sim P_X$, then*

$$-\frac{1}{n}\log P(x_1^n) \to H(\mathcal{X}) \quad \textit{in probability}$$

Theorem 2 *The typical set has almost all of the probability, and all of the sequences have about the same probability, more precisely:*

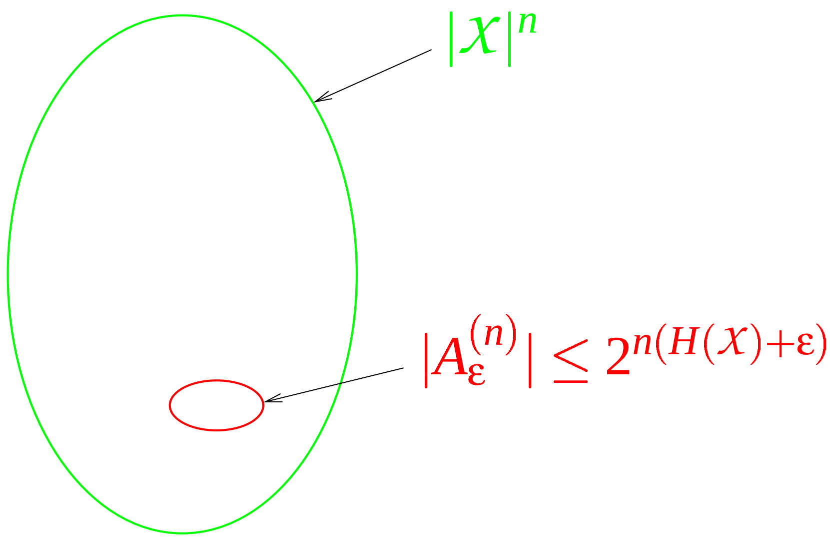
1. $x_1^n \in A_\varepsilon^{(n)} \quad \rightarrow \quad H(\mathcal{X}) - \varepsilon \leq -\frac{1}{n}\log P(x_1^n) \leq H(\mathcal{X}) + \varepsilon$
2. $Pr\{A_\varepsilon^{(n)}\} > 1 - \varepsilon$ *for n large enough.*
3. $\left|A_\varepsilon^{(n)}\right| \leq 2^{n(H(\mathcal{X})+\varepsilon)}$
4. $\left|A_\varepsilon^{(n)}\right| \geq (1-\varepsilon)2^{n(H(\mathcal{X})-\varepsilon)}$ *for n large enough.*

[a] Cover and Thomas state and prove these theorems for i.i.d. sources. They are true for ergodic sources.

Typical Set Coding

Theorem 3 *If the source X_1^n is i.i.d. and $\delta > 0$ then for n large enough there is a code which maps the source to bit strings in a one-to-one fashion and the expected value of the string length satisfies*

$$E\left[\frac{1}{n} l(X_1^n)\right] \leq H(\mathcal{X}) + \delta$$



Picture proof of Theorem 3. If you represent each $x_1^n \in A_\varepsilon^{(n)}$ by a string of $n(H(\mathcal{X})+\varepsilon)$ bits and you represent each $x_1^n \notin A_\varepsilon^{(n)}$ by a string of $n\log(\mathcal{X})$ bits, then the expected value of the number of bits used is bounded by $N_{\text{bits}} \leq n(H(\mathcal{X})+\varepsilon)+\varepsilon n\log(\mathcal{X}) = nH(\mathcal{X})+\varepsilon n(1+\log(\mathcal{X}))$. Thus the number of bits per character can be made arbitrarily close to the entropy.

Cost of Model Error

$$D(P||Q) \equiv E_P \log \frac{P(X)}{Q(X)}$$

is the *relative entropy* or Kullback Libeler distance between the two probability functions P and Q.

The average cost incurred by building an optimal code based on a *model* Q when in fact P is true is $D(P||Q)$.

Code length - probability duality

For uniquely decodablity, codeword lengths $\{l_i\}$ must satisfy

$$\sum_i 2^{-l_i} \leq 1 \tag{1}$$

Minimizing

$$L \equiv E(l(X)) = \sum_x P(x) l(x) \tag{2}$$

subject to Eqn. 1, one finds that the optimal lengths are $l^*(x) = -\log(P(x))$, and $L = -\sum_x P(x)\log(P(x)) = H(\mathcal{X})$.

If lengths are interpreted as logs of probabilities, requiring Eqn. 1 to be an equality is the same as requiring that the probabilities sum to one.

Lengths and Probabilities cont.

Conversely, one can derive a probability function Q from a set of code lengths $\{l_i\}$

$$Q(x) = \frac{2^{-l(x)}}{\sum_y 2^{-l(y)}}$$

From this point of view, Huffman coding is the solution to the problem:

- Given a probability function P, find the probability function Q that minimizes

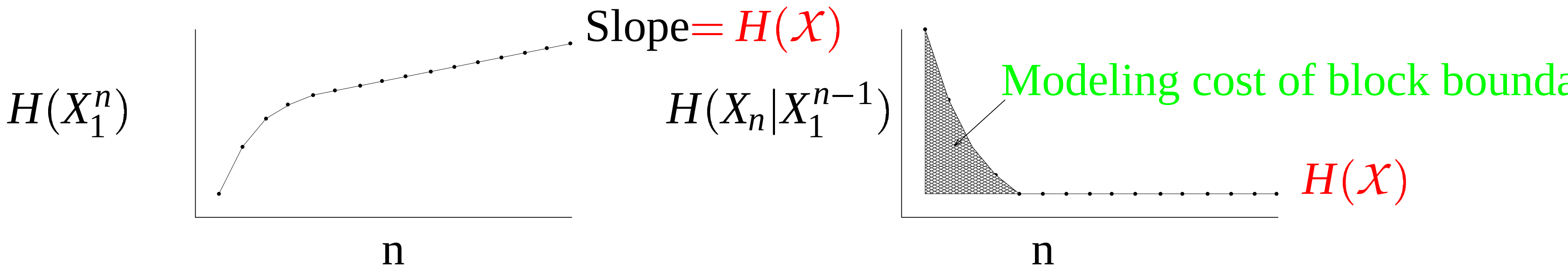
$$D(P||Q)$$

- Subject to

$$\log_2(Q(x)) \in \mathbb{Z}, \ \forall x,$$

ie, find the dyadic Q closest to P.

Costs of Blocking[a]



Plot on the left illustrates the definition of entropy rate, ie, $H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1^n)$. Plot on the right illustrates the cost of disregarding the history at block boundaries.

There are two costs of blocking:

- Modeling cost (see Fig.).
- Coding cost (can be limited to one bit per block).

---

[a] See Ross Williams' dissertation (http://www.ross.net/compression/index.html) for an informative rant on the weaknesses of block Huffman coding.

Costs of Blocking cont.

Given $P_{X(t)|X_1^{t-1}}$, i.e., a *model*, one can use arithmetic coding to represent an entire message in a single block, thus paying the blocking costs only once.

Suppose that we use blocks of size n and use the following notation:

P_{X^*} The *true* model.

$Q_{X_1^n}$ The best block model.

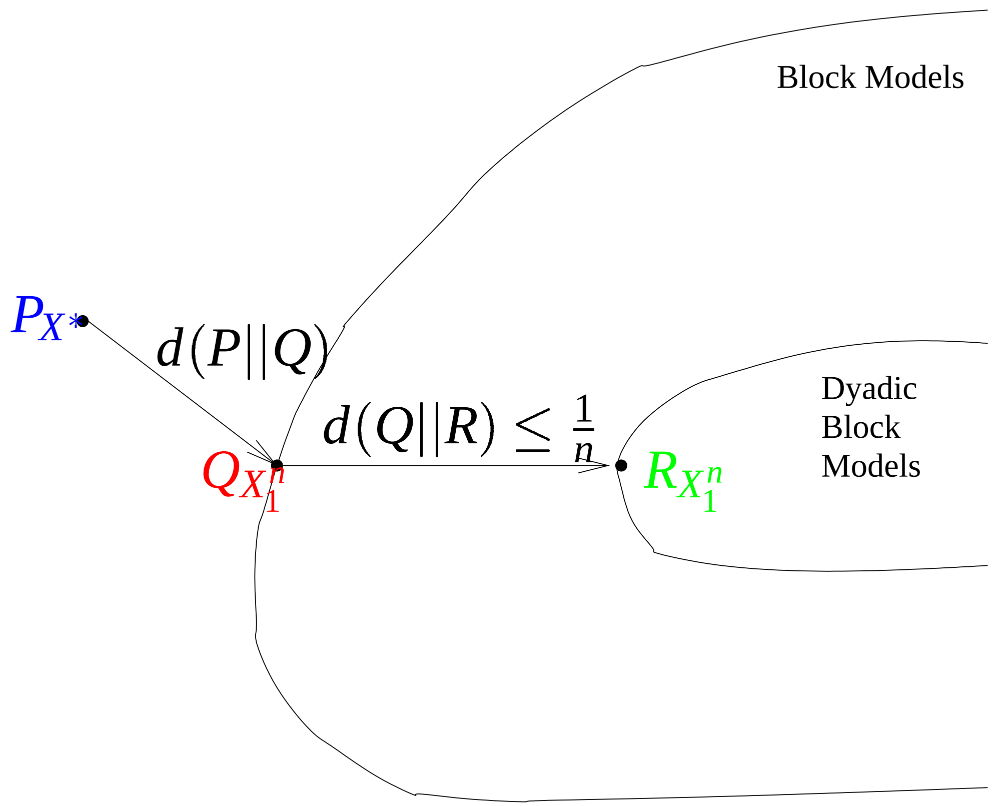
$R_{X_1^n}$ The best dyadic block model.

Asymptotically, the cost of blocking on a per symbol basis is given by the relative entropy rate, $d(P||R) \equiv \frac{1}{n} E_P \log \frac{P(X_1^n|X_{-\infty}^0)}{R(X_1^n)}$. Note

$$d(P||R) = d(P||Q) + d(Q||R)$$

as illustrated in the Fig.

Costs of Blocking cont.



Modeling Cost: $d(P||Q)$, constrain Q to be *Block Model*

Coding Cost: $d(Q||R)$, constrain R to be *Dyadic Block Model*

AR Models for image coding

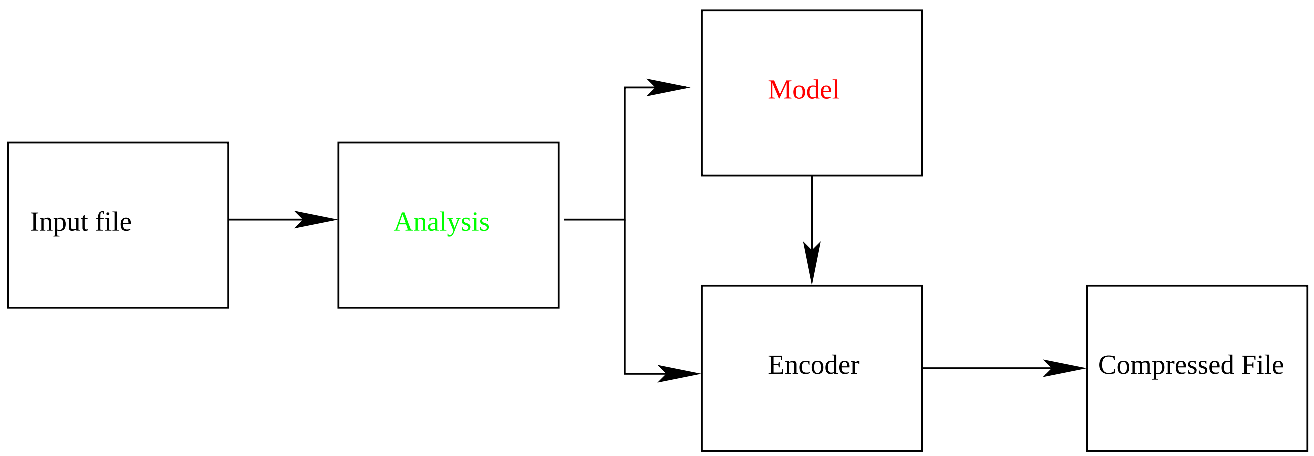
x_1	x_2
x_3	y

$\hat{y}$ the central forecast value pixel y

A the matrix of autoregressive coefficients for constructing $\hat{y}$

$$\hat{y} = \mathbf{A} \cdot \left[x_1, x_2, x_3\right]$$
$$(y - \hat{y}) \sim \mathcal{N}(0, \Sigma_y)$$

Data Flow Diagram



One can test new modeling ideas by modifying only the *Analysis* and *Model* modules.